A Novel Feature Decomposition Method to Develop Multi-hierarchy Model

Qing-Dong Wang, Hua-Ping Dai, and Youxian Sun

Abstract—**The comprehensibility of a model is very important since the results should be ultimately be interpreted by a human. This paper presents a new machine learning method, named feature decomposition method based on rough set theory, to discover concept hierarchies and develop a multi-hierarchy model of database. First the features with more relations are selected into a feature group. Then some measures by rough set theory are presented in this paper. According to these measures, the objects defined on the proposed feature group are labeled to discover a new concept. The new concept hierarchies of the database usually have specific meaning, which increase the transparency of data mining process. Finally the rule induction can process on the concept hierarchies of the database to develop a new multi-hierarchy model. The idea presented is illustrated with examples and datasets from UCI Machine Learning Repository. The results show that the multi-hierarchy model established by feature decomposition method can get high classification accuracy and have better comprehensibility.**

Keywords: rough set, feature decomposition, classification, rule induction, concept hierarchy

I. Introduction

As computer and database technologies constantly advance, the recent advance in data mining has produced many algorithms for developing classification model. The classification techniques can be implemented on variety of domains like marketing, finance and manufacturing. Fayyad et al. [1] claim that the explicit challenges for the KDD research community is to develop method that facilitate the use of data mining algorithms for real-world databases. One of the main drawbacks of many machine learning techniques is the incomprehensibility of the model produced. Traditionally most researchers pay more attention to the high performance of model, although the comprehensibility of the model is ignored.

In fact, the comprehensibility of the model is very important since the results should ultimately be interpreted by a human. The need for comprehensibility is particularly important when someone needs to be accountable for the decision, e.g. in medical decisions. For rule induction models, we often have daunts about whether they are logical, feasible, too many or too few and whether they offend common sense. In comparison with mining on raw data, Mining of transformed data sets exhibits about the same classification accuracy with the increased transparency and lower complexity of the developed models. The decomposition may take place in space and time. The most useful form of transformation of data sets is decomposition [3]. The area of decomposition in time is extensive [4]. There are two forms of decomposition in space, feature set decomposition [2,3,5] and object set decomposition [6].

In an information decision system, condition features may show various degree of relevance to the target feature, i.e. class. The relevance can be measured with the relationship of indiscernibility in rough set theory or with mutual Entropy in information theory. There are complex relationships between the conditional features. The data set used for mining is often dictated by its availability or being generated by another application [3]. For example, in industrial applications the data mining sets are often files with the statistical process control or design of experiments data. The features coming from the same sensor may have more relations. Besides concentrating on individual features, the feature set can be divided into some feature groups by intermediate concepts, which are based on aggregation of the original features. This feature relevance decomposition becomes especially important in data sets with many features. Zupan et al. [5] presented a function decomposition approach for machine learning. According to the approach, the new concept was formed based on function decomposition method, an approach originally developed to assist in the design of digital circuits [7]. But the approach is limited to consistent datasets and nominal features. Kusiak [8] introduced feature bundling method, which is of particular interest in temporal data mining as relationships are formed among features rather than their values. The relationships among features tend to be more stable in time comparing to the relationships among feature values.

Although many learning methods attempt to either extract or construct features, both theoretical analyses and experimental studies indicate that many algorithms are inscrutable to the users [10]. Moreover these methods do not

Manuscript received September 9, 2004. This work was supported by China 973 Plan under grant number 2002CB312200.

The author is with National Key Laboratory of Industry Control Technology, Zhejiang University, Hangzhou, 310027, P.R.China (e-mail: qdwang@ iipc.zju.edu.cn).

attempt to use all the relevant features and ignore the relationship between the condition attributes. In this paper, we present a new feature decomposition method to discover concept hierarchies and develop a multi-hierarchy model. This method based on rough set theory of Pawlark [9] facilitates the creation of a multi-attributes model. The features with high relevance can be selected by skill and experience of expert and aggregate into one feature group. Then information gain is computed to measure the significance of feature groups and get rid of the redundant features. According to some measures of rough set theory, the objects defined on the proposed feature group are labeled by a new intermediate concept. The concept hierarchies of the database have specific meaning, which increased the transparency of data mining process and enhance the comprehensibility of the model. Each feature group and the corresponding intermediate concept compose the subset of the database. Attributes reduct and rule induction can process on the subsets. Finally rule induction can be processed by rough set theory, which can reduce rule set with the intermediate concept even the dataset contain uncertainty or noise information.

The rest of the paper is organized as follows. Section 2 introduces some basic concept of Rough Set Theory. Section 3 presents the feature decomposition algorithm in detail. Section 4 the method is experimentally evaluated on several dataset coming from UCI depository and the last section contains final conclusions.

II. Preliminaries

The rough set methodology was introduced by Pawlak [9] in the early 1980s as a mathematical tool to deal with uncertainty. Here, we only introduce the basic notation from rough set approach used in the paper. The main data set in this paper is a decision table. Formally, the decision table is defined as 4-tuple *DT=(U, A, V, f)*, where *U* is a non-empty finite set of objects and *A* is a non-empty finite set of condition *C* and decision *D* attributes, such that $C \bigcup D = A$ and $C \bigcap D = \varnothing$. *V* is a non-empty finite set of attribute values. Each attribute $a \in A$ can be viewed as a function that maps elements of *U* into a set V_a. *f* is an information function, $f : U \times A \to V$.

Let $IND(P) \in U \times U$ denote an indiscernibility relation defined for a non-empty set of attributes $P \subseteq A$ as:

$$IND(P) = \left\{ (x, y) \in U \times U : \underset{q \in P}{\forall} f(x, q) = f(y, q) \right\}$$

If $(x, y) \in IND(P)$ we will say that x and y are *P-indiscernible*. Equivalence classes of the relation $IND(P)$ are referred to as *P-elementary sets*. In the rough set approach the elementary sets are the basic building blocks of knowledge. The family of all equivalence classes of $IND(P)$, i.e., the partition determined by *P*, will be denoted by *U/P*.

For every subset $X \subseteq U$ we define the lower approximation $B_*(X)$ and the upper approximation $B^*(X)$ as follows:

$$B_*(X) = \left\{ x \in U : [x]_B \subseteq X \right\},$$

$$B^*(X) = \left\{ x \in U : [x]_B \cap X \neq \varnothing \right\}.$$

Another important issue in data analysis is discovering dependencies between attributes. Let *D* and *C* be subsets of *A*. we will say that *D* depends on *C* in a degree k, denoted $C \Rightarrow_k D$, if $k = \gamma(C, D) = \dfrac{|POS_C(D)|}{|U|}$

Where $POS_C(D) = \bigcup_{X \in U/D} C_*(X)$ called a positive region of the partition *U/D* with respect to *C*, is the set of all elements of *U* that can be uniquely classified to blocks of the partition *U/D*, by means of *C*. The coefficient *k* expresses the ratio of all elements of the universe, which can be properly classified to blocks of the partition *U/D*, employing attributes *C* and will be called *the degree of the dependency*.

III. Feature Decomposition Algorithm

In this section, we first present the two measures by rough set theory for discovering the intermediate concept. Information gain is computed to measure the significance of feature groups and get rid of the redundant features. After that, the algorithm and one example are presented to express the process of feature decomposition and model development

A. Measures based on rough set theory

Feature decomposition method attempt to select the features by some criteria to form a feature group and label the union by a new intermediate concept. About the model of the transformed dataset, the classification accuracy does not decrease significantly, and the comprehensibility will be increased. Our algorithm is a heuristic method with rough set measures to find the optimal partition of the intermediate concept c_i. The following are the measures based on rough set theory:

1) Consistency measure: Given the feature group G_i and the corresponding intermediate concept c_i , We define a criteria based on the degree of dependency for partition evaluation as below.

$$J_{c1}=\left|\frac{\gamma(C,D)-\gamma\left(R\bigcup\{c_i\},D\right)}{\gamma(C,D)}\right|\le\delta$$

where $\gamma\left\{R\bigcup\{c_i\},D\right\}$ expresses the degree of dependency between attributes $R\bigcup\{c_i\}$ and *D*, *R* means the remains features except feature group G_i, δ is a user given threshold. This definition expresses that $DT_{new}=\left\{U,R\bigcup\{c_i\}\bigcup D,V,f\right\}$ has the approximate degree of the dependency compared with *DT*. The threshold δ suits the characteristic of consistency measure because real-world data is usually noisy and if δ is set to 0 strictly then it may happen that "good" features are filtered away. Especially, when $J_{c1}=0$, the new decision system has the same classification performance as the original datasets.

2) Min-value measure: Minimizing the cardinality of V_{c_i} , represented as $card(V_{c_i})$ which expresses the number of values of the intermediate concept c_i . The smaller $card(V_{c_i})$, the simpler partition will get. Generally, the number *n* of partitions divided by the proposed feature group is limited:

$$n\le\prod_{i=1}^{k}card(a_i)$$

where a_i is the feature which is selected into the feature group. So the time complexity of discovering a intermediate concept is $O(Nm)$, where *N* is the number of objects and *m* is the cardinality of feature domains.

B. Measuring the significance of feature groups

After the selection of the feature groups, the feature set is divided into *k* feature groups $G_1, G_2,..., G_k$.. Each feature group has different significance to the decision features *D*. Some of feature groups are redundant. In order to keep the same classification performance and model transparency as the original data set, the feature group with high significance should be selected firstly to compute the corresponding intermediate concept.

Information gain has been an effective method to measure the importance of attributes [11]. So, it is feasible to measure the significance of feature groups with information-theoretic measures. Suppose the pattern class label $C_i\,, i=1,\cdots,card\left(V_d\right)$ is the corresponding value of the decision attribute *d* which divide universe *U* into partitions Θ , where $card\left(\cdot\right)$ denotes cardinality of set, $card\left(\cdot\right)$ denotes the number of decision classes, V_d denotes the value set of decision attribute *d*. If the objects randomly distribute in the pattern classes of Θ , we can define the information measure on Θ as follows:

$$H\left(\Theta\right)=-\sum_{i=1}^{card\left(V_d\right)}p\left(C_i\right)\log_2 p\left(C_i\right),$$

where $p\left(C_i\right)=\dfrac{card(C_i)}{card(U)}$ denotes the possibility that the objects on *U* is right classified.

For feature group G_i, the conditional entropy can be defined as follows:

$$H\left(\Theta\mid G_i\right)=-\sum_{j=1}^{card\left(V_{Gi}\right)}p(v_j)\sum_{i=1}^{card\left(V_d\right)}p\left(C_i\mid v_j\right)\log_2 p\left(C_i\mid v_j\right).$$

TABLE Ⅰ
DATA TABLE

No.	outlook	temperature	humidity	class
1	1	1	1	N
2	2	1	1	P
3	3	2	1	N
4	3	1	2	P
5	3	3	2	N
6	2	3	2	P
7	1	2	1	N
8	1	3	1	P
9	1	3	2	P
10	3	2	2	N
11	1	2	2	P
12	2	2	1	P
13	2	1	2	P

TABLE Ⅱ
THE RULE SET OF SUBSET

Rule	Coverage

{temperature=1,humidity=1}
→{c1=1}
	15.4%

{temperature=2,humidity=1}
→{c1=1}
	23.1%

{temperature=1,humidity=2}
→{c1=2}
	15.4%

{temperature=2,humidity=2}
→{c1=3}
	15.4%

{temperature=3}
→{c1=3}
	30.8%

TABLE Ⅲ
THE RULE SET OF SUBSET

Rule	Coverage
{outlook=1,c1=1}→{class=0}	15.4%
{outlook=3,c1=1}→{class=0}	7.7%
{outlook=3,c1=3}→{class=0}	15.4%
{outlook=1,c1=3}→{class=1}	23.1%
{outlook=2}→{class=1}	30.8%
{c1=2}→{class=1}	15.4%

The mutual entropy of condition attribute G_i and decision attribute d can be defined as

$$I\left(\Theta,G_i\right)=H\left(\Theta\right)-H\left(\Theta\mid G_i\right)$$

$$=\sum_{j=1}^{card(V_{Gi})} p(v_j)\sum_{i=1}^{card(V_d)} p\left(C_i\mid v_j\right)\log_2 p\left(C_i\mid v_j\right)-\sum_{i=1}^{card(V_d)} p\left(C_i\right)\log_2 p\left(C_i\right)$$

The larger the value $I\left(\Theta,G_i\right)$, the stronger the relationship between condition attribute G_i and target attribute d, and the more important is the attribute G_i.

C. Description of the algorithm

Algorithm. Feature decomposition method using rough sets
Given: A *N*-case data set T containing *n*-dimensional patterns, labeled by *l* associated classes, denoted by $DT=(U, C\cup D)$.

1. Select the features in terms of expert's experience to form some feature groups $G_1, G_2,\ldots, G_k$.
2. Compute the significance of feature groups by the information theory.
3. According to the significance sequence of feature groups, compute the intermediate features $c_1, c_2, \ldots, c_k$ by the two measures of rough set theory
4. Attribute reduct, induce the rule sets by rough set theory on data set $DT_i=(U, G_i \cup c_i), i=1,2\ldots,k$

Induce the rule set by rough set theory on data set $DT'=(U, \{c_1,c_2,\ldots,c_k\} \cup D)$ to obtain the multi-hierarchy model.

Example. The example presented next illustrates the decomposition process. Consider the decision table in Table1 containing 13 objects. Each object is described by three condition features, *outlook*, *temperature* and *humidity*. According to the classic definition, the feature decomposition is to divide the condition feature set into several sub-feature sets. For example, the condition feature set can be decomposed into R_1 and R_2, $R_1 \cap R_2 = \varnothing$, $R_1 \cup R_2 = C$. The features in each group have high relevance relationship. For the data set in table Ⅰ, $R_1=\{outlook\}$, $R_2=\{temperature, humidity\}$. We could give the new label c_1 to the objects according to the relationship of indiscernibility defined on subset, e.g. R_2. We can call c_1 the comfort index by experience. From the view of rough set theory, the set *U* is partitioned by the relation of indiscerniblity defined on each feature group as below.

$$U/R_1=\{(1,7,8,9,11),(2,6,12,13),(3,4,5,10)\}$$

$$U/R_2=\{(1,2),(3,7,12),(4,13),(5,6,9),(8),(10,11)\}$$

$$U/class=\{(1,3,5,7,10),(2,4,6,8,9,11,12,13)\}$$

According to the measures above, the intermediate concept is discovered. The rule sets of subset (see table Ⅱ) and the transformed dataset (see table Ⅲ) are listed below:

IV. Experiments

The data sets used in the feature decomposition experiments are real-life data sets coming from ***UCI Machine Learning Repository*** [12]. The experiment was conducted on Monk's problems [13] at first to check the usefulness of the rule sets generated by our algorithm in terms of their predictive accuracy, and the experiment was executed with *holdout* technology. The MONK's problems rely on an artificial robot domain, in which robots are described by six different attributes. The MONK's problem includes three sub-problems, and each problem is given by a logical description of a class. Detailed descriptions of Monk's problems can be found in [12].

TABLE Ⅳ
COMPARATIVE RESULTS OF CLASSIFICATION ACCURACY (%)

Algorithms	M_1	M_2	M_3
Naïve Bayesian[12]	70.8	67.1	97.2
ID3[12]	98.6	67.9	94.4
C4.5[12]	100	64.8	94.4
RSES[12]	88.7	73.6	94.7
LERS[12]	100	84.4	94.1
Our algorithm	100	89.5	85.8

The training set of MONK1 (M_1), MONK2 (M2) and MONK3 (M3) contains 124, 169 and 122 objects separately. The test sets all have 432 objects. M1 and M2 have no noise data, but M3 has 5% misclassifications, i.e. noise in the training set. In the six attributes $x_1, x_2,\ldots, x_6,$, x_1 means head shape, and x_2 means body shape. So we can define a new intermediate concept c_1 which means shape factor of the objects by the experience. On each of the MONK's problem, our algorithm is presented with training set and examined the predictive accuracy of the induced rules on a test set. Table 4 reports the testing accuracy of our algorithm on each MONK's problem in comparison with other well-known machine learning algorithm.

Table Ⅳ presents the experiment results of our algorithm and some typical machine learning algorithm, Naïve Bayesian, ID3, C4.5, RSES and LERS method on three MONK's problems. It can be seen that they have respective advantage to settle with MONK's problems. M1 is a simple classification problem, and most of the algorithm can obtain preferable accuracy. And our algorithm can give 100% accurate prediction as many of other algorithms. For M2, because of the complexity of the features combination, most of the algorithm cannot produce excellent classification rules. But the accuracy obtained by our algorithm on M2 reaches to 89.5%, which is obviously higher than the classification accuracy of other algorithms. The reason for increased classification accuracy with the intermediate concept might due to the fact the associations among features and decisions are stronger than those built on single feature. In fact, the intermediate concept can be looked as the target of the regression function defined on a subset of features.

The information systems on problem M1 and M2 are consistent datasets. But problem M3 has noisy data, i.e. the information system is an inconsistent dataset. Our algorithm can handle the problem and discover the new concept from the noisy data. From table 3, it can be seen that the accuracy for our algorithm can achieve 85.8% and the result is inferior to other algorithm to a certain extent. The existence of noisy data decreases the performance of classification model. Nevertheless our algorithm can develop a multi-hierarchy model with high comprehensibility. The new concept which has specific meaning can be discovered by our algorithm. The value number of the new concept c1 on M1, M2 and M3 is 2, 3 and 5 respectively. For example, the two values of the intermediate concept on M1 are easy to interpret. If the body shape is the same as the head shape, the intermediate concept is equal to 1. If not, it is equal to 2. And our algorithm maps 9 combinations of attribute_1 and attribute_2 to only two values of the intermediate concept. In a word, the multi-hierarchy model developed by our algorithm is not only a powerful model with high performance but also a transparent model with better comprehensibility.

V. Conclusions and future work

In the most data mining applications, the model is developed on the original data set. In this paper a new method named feature decomposition algorithm was introduced. Two measures based on rough set theory are presented to discover the new intermediate concept. And rule induction is processed by rough set theory to build the multi-hierarchy model. Experiments on MONK's problems are made to evaluate the performance of the method. The results show that the multi-hierarchy model established by feature decomposition method can get high classification accuracy and have better comprehensibility. We conclude the main advantages of the feature decomposition method as follows:

1. This method can enhance the transparency of the model. The model developed by the feature decomposition method expresses the hierarchy structure of database clearly. And the rule sets induced by the method are comprehensible. So in decision support tasks, reasons for the decision can be clearly identifiable.
2. This method can deal with the inconsistent dataset, process appropriate treatment of noise in data. Though the existence of noise degrades the performance of model, the results of experiment verify that this method can build multi-hierarchy model with high comprehensibility from the noisy data.

The future works are as follows. When dealing with the databases without any background knowledge, what criterion we conform to form the feature groups. It is therefore desirable to find some new criterion to handle this problem. The approach of statistics maybe can use for reference, such as *factor analysis technology*. Another interesting issue is to decrease the effect of the noisy data. It will be useful to enhance performance of the model by combining feature decomposition method and the approaches of handling noise.

References

[1] Fayyad, U., Piatesky-Shapiro, G., and Smyth P., *From Data Mining to Knowledge Discovery: An Overview*, Advances in Knowledge Discovery and Data Mining, pp1-30, MIT Press, 1996

[2] Maimon, O., and Rokach L., *Improving Supervised Learning by Feature Decomposition*, FoIK 2002, LNCS 2284, pp.178-196, 2002

[3] Kusiak, A., *Decomposition in Data Mining: An Industrial Case Study*, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 23, No. 4, 2000, pp.345-353

[4] M.J. Zaki and C.T.Ho, Eds., *Large-Sclae Parallel Data Mining*. New York: Springer-Verlag, 2000

[5] Zupan, B., Bohanec, M., Demsar, J., and Bratko, I., *Feature Transformation by Function Decomposition*, IEEE intelligent systems & their applications, 13:38-43, 1998

[6] Chan, P.K. and Stolfo, S.J., *A Comparative Evaluation of Voting and Meta-learning on Partitioned Data*, Proc. 12th Intl. Conf. on Machine Learning ICML-95, 1995

[7] H.A. Curtis, *Anew Approach to the Design of Switching Functions*, Van Nostrand, Princeton, 1962

[8] Kusiak, A., *Feature Transformation Methods in Data Mining*, IEEE Transactions on Electronics Packaging Manufacturing, Vol. 24, No. 3, 2001, pp.214-221

[9] Pawlak Z. *Rough Sets. Theoretical Aspects of Reasoning About Data*, Kluwer Academic Publishers, Dordrecht, 1991

[10] Ridgeway, G., Madigan, D., Richardson, T. and O'Kane, J., *Interpretable Boosted Naïve Bayes Classificatioin*, Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pp 101-104

[11] Yao Y.Y, Wong S.K.M, Butz C.J, On information-theoretic measures of attribute importance. 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining (PKDD'99), 133-137,1999

[12] UCI repository of machine learning databases (1996). http://www.ics.uci.edu/~mlearn/mlrepository.html. Department of information and computer science, university of California

[13] S. B. Thrun, J. Bala, E. Bloedorn, I. Bratko, e tal. The MONK's problems: a performance comparison of different learning algorithms. Technical Reports. Carnegie Mellon University. CMU-CS-91-197. December 1991